

•
•
•
•
•
•
•
•

Asynchronous Zero-copy Communication for Synchronous Sockets in the Sockets Direct Protocol over InfiniBand



P. Balaji, S. Bhagvat, H. -W. Jin and D. K. Panda

Network Based Computing Laboratory (NBCL)

Computer Science and Engineering

Ohio State University

• • • • • • • •

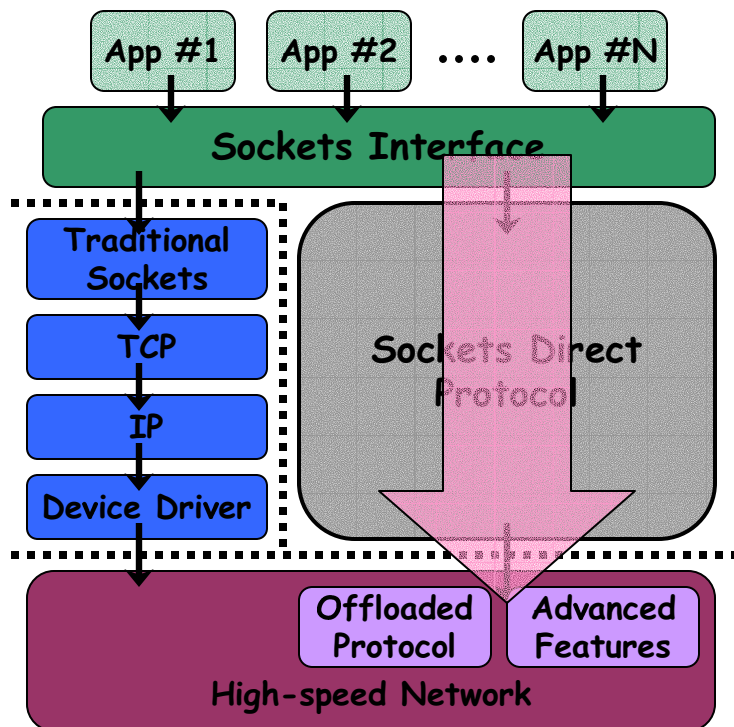


InfiniBand Overview



- An emerging industry standard
- High Performance
 - Low latency (about 2us)
 - High Throughput (8Gbps, 16Gbps and higher)
- Advanced Features
 - Hardware offloaded protocol stack
 - Kernel bypass - direct access to network for applications
 - RDMA operations - direct access to remote memory

Sockets Direct Protocol (SDP)



- High-Performance Alternative to TCP/IP sockets for IB, etc.
- Hijack and redirect socket calls
- Application transparent
 - Binary compatibility (most cases)
- Utilizes IB capabilities
 - Offloaded Protocol
 - RDMA operations
 - Kernel bypass

Sockets APIs Supported by SDP

	Synchronous Sockets	Asynchronous Sockets	Extended Sockets (OSU Specific)*
Communication	Synchronous	Asynchronous	Asynchronous
Operations Outstanding	At most one	More than one	More than one
SDP Implementations	BSDP, ZSDP, AZ-SDP	BSDP, ZSDP	BSDP, ZSDP
Existing Applications	Most	Few	Very few
Potential for Performance	Limited	High	High

(Portions of this table have been borrowed from Mellanox Technologies)

* RAIT05: "Supporting iWARP compatibility and features for regular network adapters". P. Balaji, H. -W. Jin, K. Vaidyanathan and D. K. Panda. RAIT Workshop; in conjunction with Cluster '05

04/25/06

Pavan Balaji (The Ohio State University)

Presentation Layout

§ Introduction and Background

§ Understanding Asynchronous Zero-copy SDP

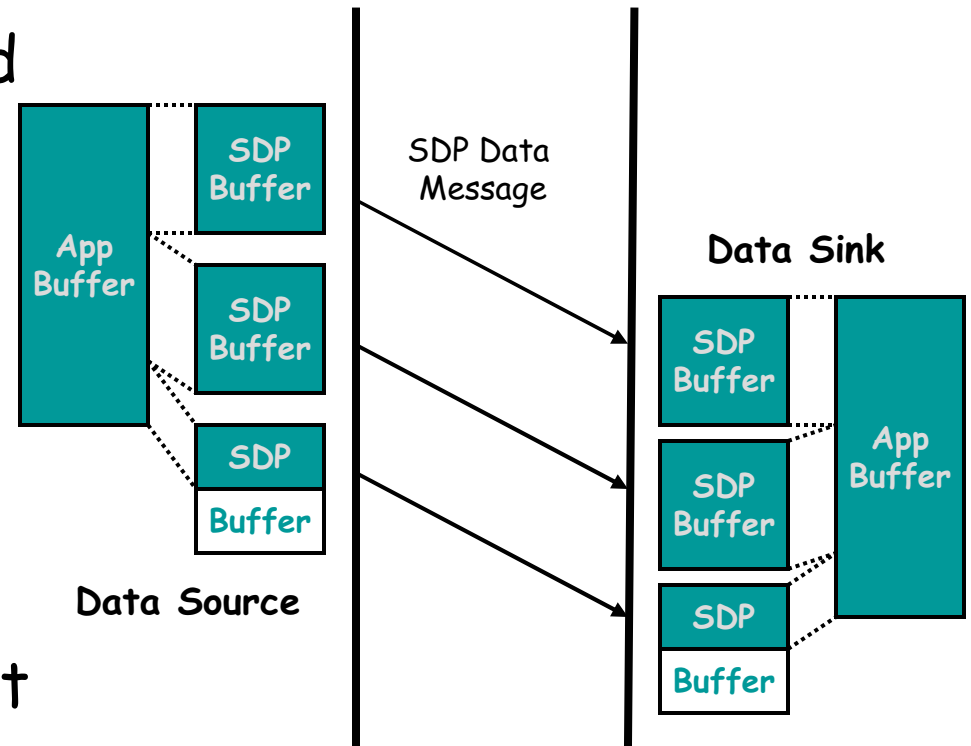
§ Design Issues in AZ-SDP

§ Performance Evaluation

§ Conclusions and Future Work

Buffer-copy SDP (BSDP)

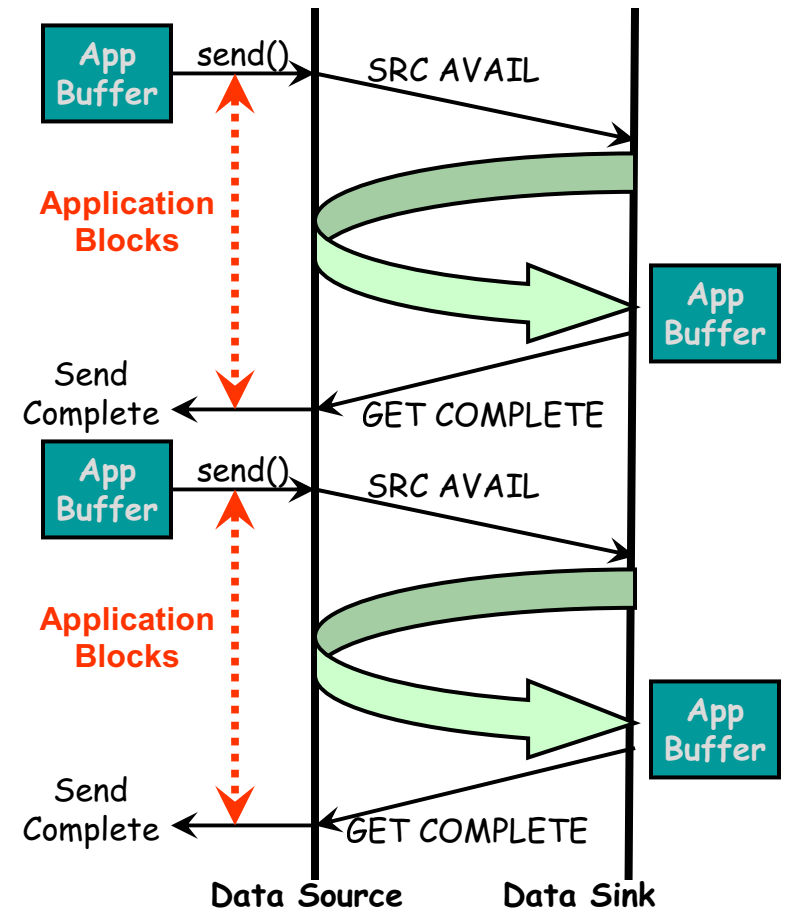
- Several buffer-copy based implementations of SDP exist
 - OSU, Mellanox, Voltaire
- HCA offloads transport and network layers
- Copy overhead still present



ISPASS04: "Sockets Direct Protocol over InfiniBand in Clusters: Is it Beneficial?". P. Balaji, S. Narravula, K. Vaidyanathan, S. Krishnamoorthy and D. K. Panda. IEEE International Conference on Performance Analysis of Systems and Software (ISPASS), 2004.

Zero-copy SDP (ZSDP)

- Implemented by Mellanox
 - RDMA Read based design
- Benefits of zero-copy
- Limited by the API of Synchronous Sockets
 - At most one outstanding communication request
 - Control message latency (50% time for 16K message)
- Intolerant to Skew

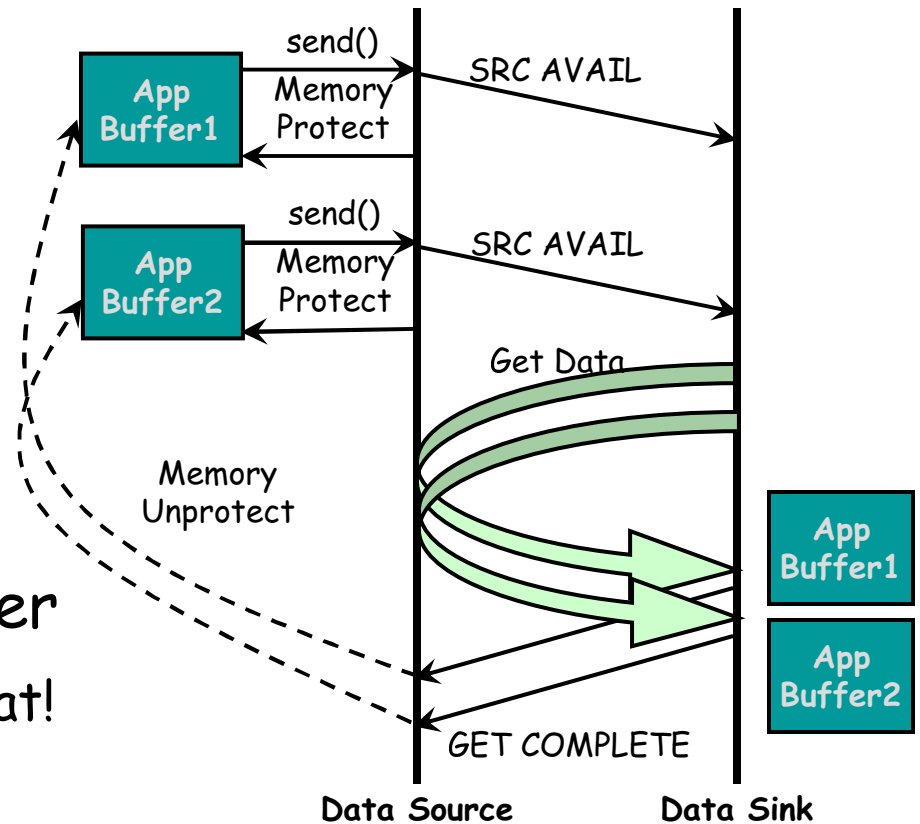


Asynchronous Zero-copy SDP (AZ-SDP)

- Basic zero-copy communication is synchronous
 - Data communication accompanied by control messages
 - Communication will be latency bound
- Asynchronous Zero-copy SDP
 - Utilize the benefits of asynchronous communication (more than one outstanding communication operation)
 - Maintain the semantics of synchronous sockets (application can assume that it is using synchronous sockets)
 - Objectives: Correctness, Transparency and Performance
 - Key Idea: Memory protect buffers

AZ-SDP Functionality

- Send returns as soon as communication is initiated
 - Application "thinks" communication is synchronous
- Memory unprotected after communication completes
- If application touches buffer
 - Communication complete: Great!
 - Else PAGE FAULT generated





Presentation Layout



§ Introduction and Background

§ Understanding Asynchronous Zero-copy SDP

§ Design Issues in AZ-SDP

§ Performance Evaluation

§ Conclusions and Future Work

Design Issues in AZ-SDP

Handling a Page Fault

- Block-on-Write: Wait till the communication has finished
- Copy-on-Write: Copy data to internal buffer and carry on communication
- Handling Buffer Sharing
 - Buffers shared through mmap()
- Handling Unaligned Buffers
 - Memory protection is only in the granularity of a page
 - Malloc hook overheads



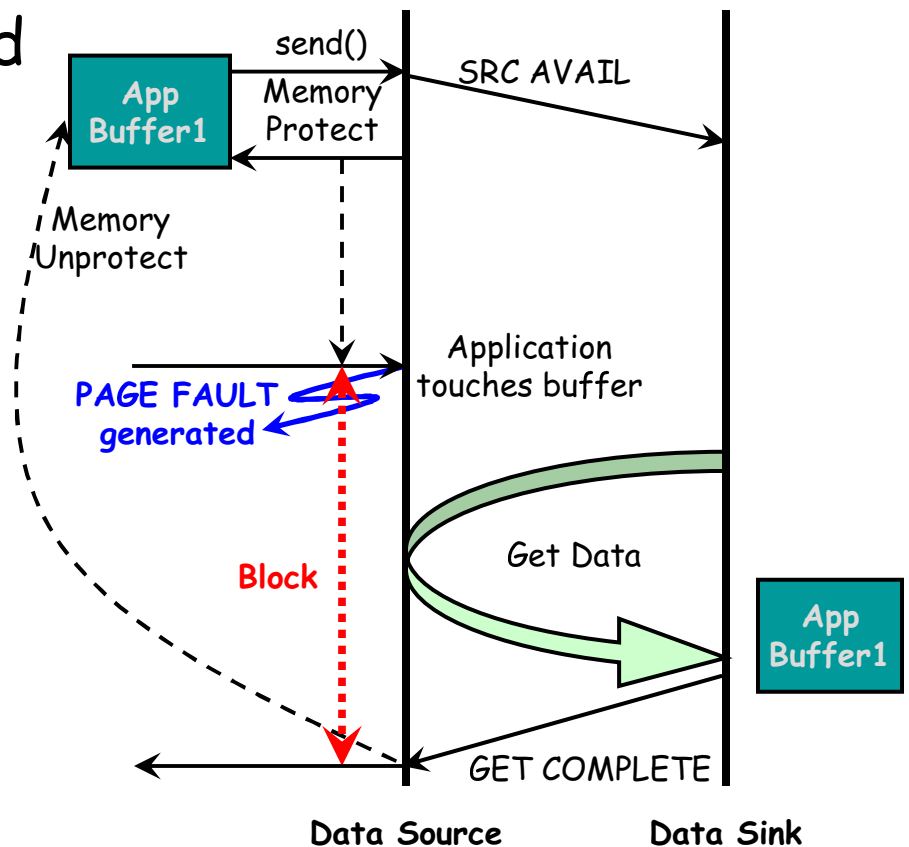
Handling a Page Fault



- Memory protection needed to disallow the application from accessing an occupied communication buffer
- Page fault generated on access
 - Number of page faults generated are application dependent
- Two approaches for handling the page-fault
 - Block on Write
 - Copy on Write

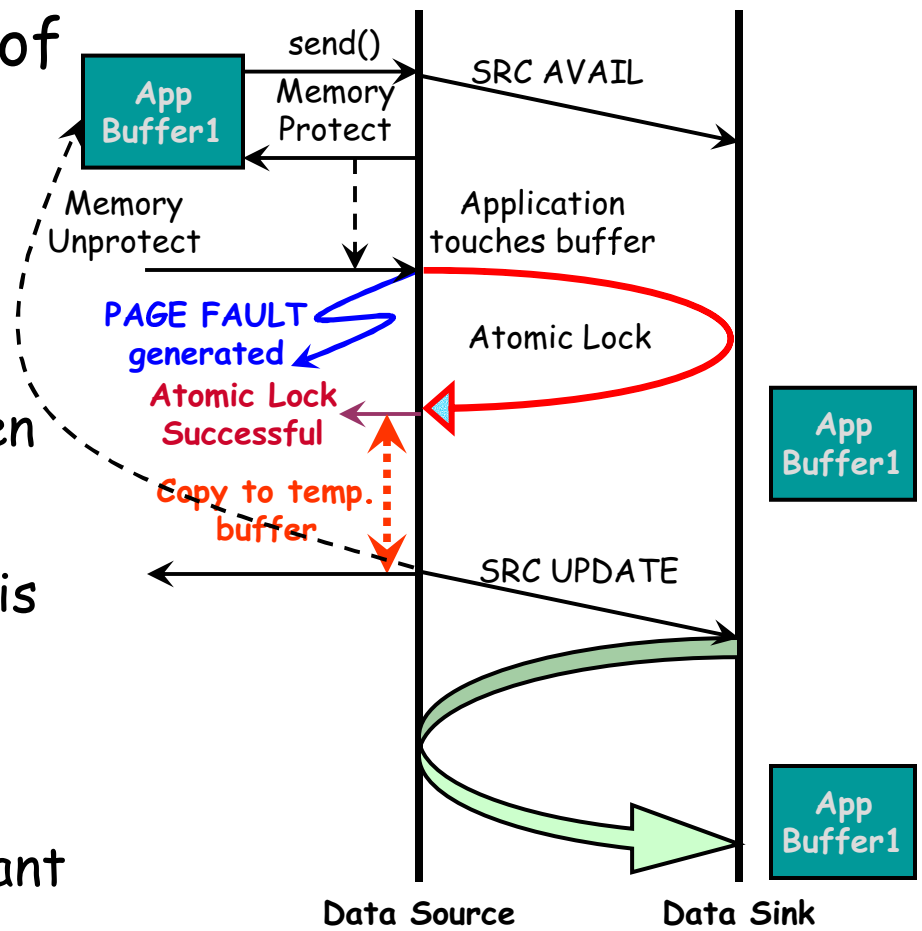
Block-on-Write

- Optimistic approach to avoid blocking for communication
 - ZSDP blocks during the communication call
 - AZ-SDP delays blocking
- Advantage:
 - Zero-copy communication
 - SDP specification compliant
- Disadvantage:
 - Not skew tolerant



Copy-on-Write

- Enhances the functionality of Block-on-Write
 - Does not blindly block
- Advantage:
 - Zero-copy communication when possible
 - Skew tolerant when receiver is not ready
- Disadvantage
 - Not SDP specification compliant





Presentation Layout



§ Introduction and Background

§ Understanding Asynchronous Zero-copy SDP

§ Design Issues in AZ-SDP

§ Performance Evaluation

§ Conclusions and Future Work

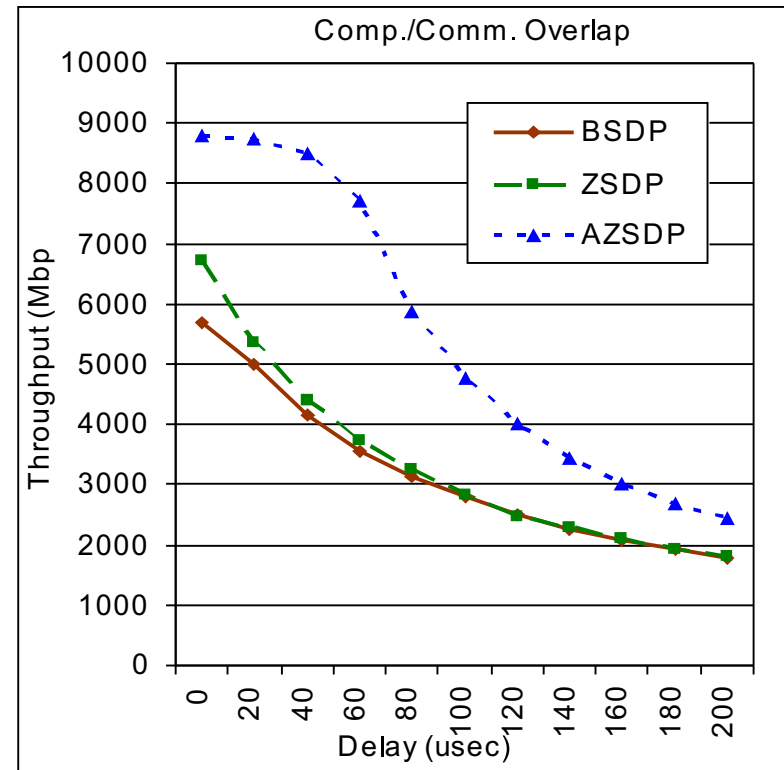
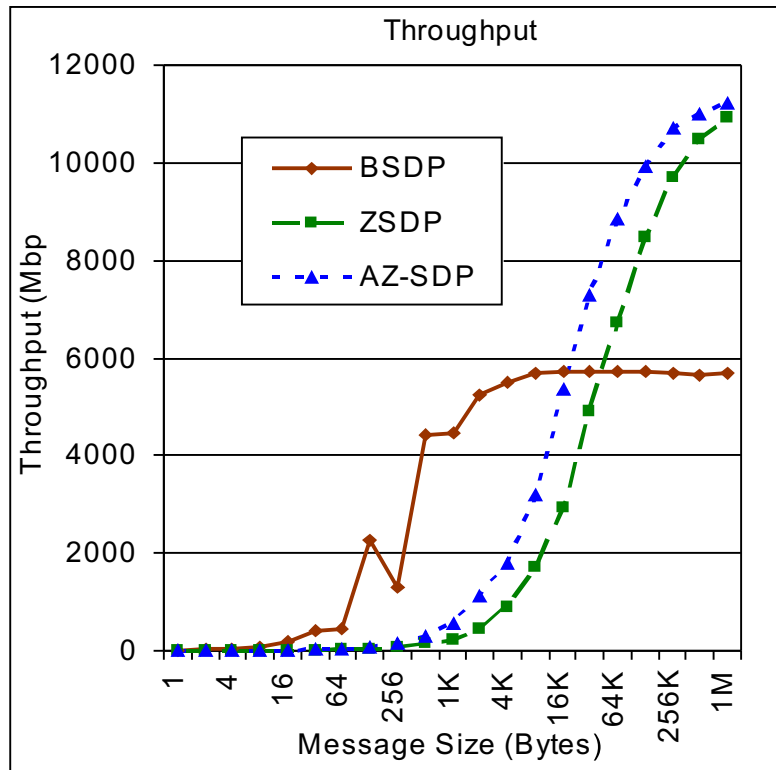


Experimental Test-Bed



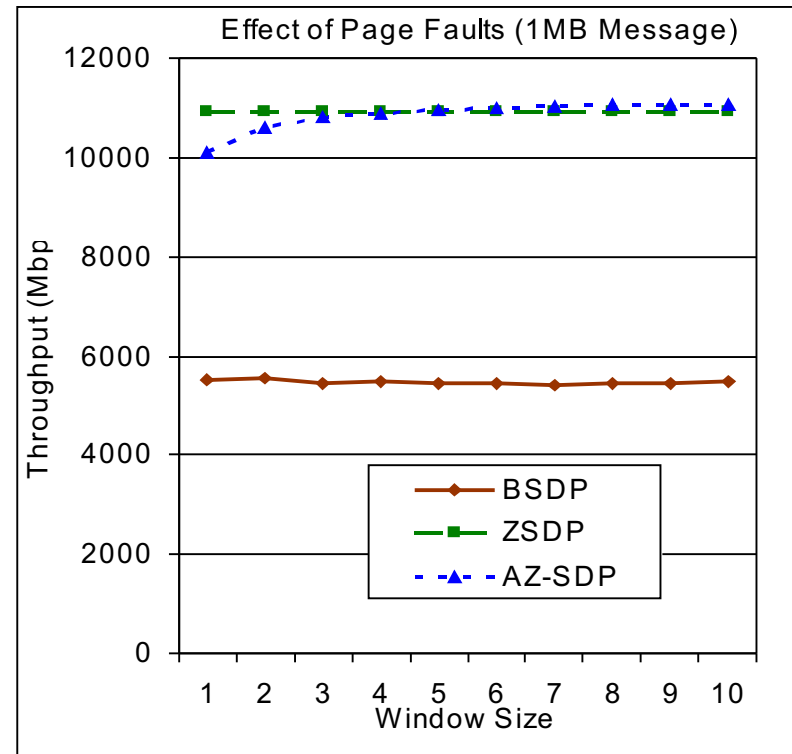
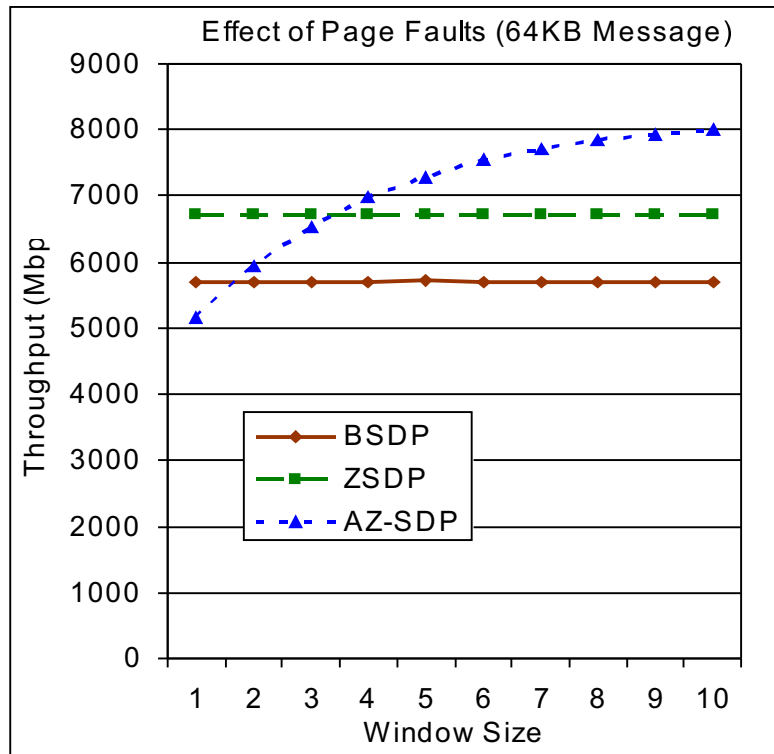
- 4 node cluster
 - Dual 3.6 GHz Intel Xeon EM64T processors (2 MB L2 cache), 512 MB of 333 MHz DDR SDRAM
 - Mellanox MT25208 InfiniHost III DDR PCI-Express adapters (capable of a link-rate of 16 Gbps)
 - Mellanox MTS-2400, 24-port fully non-blocking DDR switch

Throughput and Comp./Comm. Overlap



- 30% improvement in the throughput
- Up to 2X improvement in computation/communication overlap tests

Impact of Page-faults



- When application touches the communication buffer very frequently, PAGE FAULT overheads degrade AZ-SDP's performance



Presentation Layout



§ Introduction and Background

§ Understanding Asynchronous Zero-copy SDP

§ Design Issues in AZ-SDP

§ Performance Evaluation

§ Conclusions and Future Work

•
•
•

Conclusions and Future Work

- Current Zero-copy SDP approaches: Very restrictive
- AZ-SDP brings the benefits of asynchronous sockets to synchronous sockets in a **TRANSPARENT** manner
- 30% better throughput and 2X improvement in computation-communication overlap tests
- Analysis with applications and large-scale clusters
- Integration with OpenIB/Gen2

Acknowledgements

Our research is supported by the following organizations

- Current Funding support by



- Current Equipment support by



•
•
•

Web Pointers



NBCL

Website: <http://www.cse.ohio-state.edu/~balaji>

Group Homepage: <http://nowlab.cse.ohio-state.edu>

Email: balaji@cse.ohio-state.edu